A New Network Slicing Framework for Multi-Tenant Heterogeneous Cloud Radio Access Networks

Ying Loong Lee*, Jonathan Loo† and Teong Chee Chuah*
*Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
Email: lee.ying.loong12@student.mmu.edu.my, tcchuah@mmu.edu.my
†School of Science and Technology, Middlesex University, London NW4 4BT, United Kingdom
Email: J.Loo@mdx.ac.uk

***Abstract*—Research on network slicing for multi-tenant heterogeneous cloud radio access networks (H-CRANs) is still in its infancy. In this paper, we redefine network slicing and propose a new network slicing framework for multi-tenant H-CRANs. In particular, the network slicing process is formulated as a weighted throughput maximization problem that involves sharing of computational resources, fronthaul capacity, physical remote radio heads and radio resources. The problem is then jointly solved using a sub-optimal greedy approach and a dual decomposition method. Simulation results demonstrate that the framework can flexibly scale the throughput performance of multiple tenants according to the user priority weights associated with the tenants.**

I. Introduction

In the past decade, the demand for broadband multimedia services has been increasing explosively. With this ongoing trend, the revenue of mobile network operators will soon be exceeded by the capital expenditure (CAPEX) and the operating expenditure (OPEX) required to operate the infrastructure. In view of this predicament, several radio access network (RAN) architectures have been proposed for the next generation mobile cellular networks. Particularly, the cloud-RANs (C-RANs) have attracted much interest from academia and industry.

In a C-RAN, the baseband units (BBUs) of all remote radio heads (RRHs) are centralized at a single BBU pool via optical fronthaul links. In this architecture, the upper-layer baseband functions are carried out by the BBU pool whereas the RRHs performs radio transmission to the users. The C-RAN architecture can achieve significant reduction in energy consumption and higher throughput while providing flexibility and scalability in network deployment [1]. All of these features lead to reductions in the CAPEX and OPEX. The C-RAN architecture has further been adapted to heterogeneous cellular systems, leading to the heterogeneous C-RAN (H-CRAN) architecture consisting of a macro base station (MBS) and low-power RRHs (small cells) [2]. This architecture relieves the control signaling burden of the RRHs and delivers higher spectral and energy efficiency to the network. Currently, intensive research has been carried out for developing H-CRANs.

While C-RANs are being intensively studied, multi-tenancy has recently emerged as an interesting concept for a C-RAN shared by multiple mobile virtual network operators (VNOs) who do not own the RAN infrastructure. The basic idea of multi-tenancy is to allow multiple VNOs to share a physical RAN infrastructure. In fact, this RAN sharing can be viewed as a RAN-as-a-service (RANaaS) business model and the VNOs are known as the tenants. The idea of multi-tenancy is equivalent to multi-operator RAN sharing [3]–[5]. However, the implementation of infrastructure sharing in C-RANs require network and resource virtualization [6], [7]. In the context of network virtualization, a virtual network created for a tenant is called a *network slice* which consists of a set of data flows sharing a physical RAN with multiple other virtual networks. Currently, development in this area is still in its infancy, especially the implementation of multi-tenancy in C-RANs, thus forming the core of study of this paper. In the following, we provide a background review of multi-operator RAN sharing and multi-tenancy.

Multi-operator RAN sharing based on virtualization has been studied in [3]–[5], [8]–[10]. In [3], capacity allocation between multiple VNOs has been studied using a stochastic sequential auction game. A unique Nash equilibrium has been proved to exist as the solution to the game by introducing conjectural prices to represent future congestions of the users. In [4], the network virtualization substrate, which is composed of a slice scheduler and a flow scheduler, has been designed for virtualization of wireless resources in cellular networks. The slice scheduler isolates and reserves resources for multiple network slices whereas the flow scheduler allocates resources to the data flows of each network slice for data transmission. Comparison between physical and virtual resource sharing has been carried out in [5] and it is shown that virtual resource sharing provides better performance at the cost of complexity. In [8], a RAN multi-tenant cell slicing controller (RMSC) is devised to manage resources and balance loads among the network slices using a utility optimization approach. Resource distribution among base stations (BSs) in multi-tenant heterogeneous cellular networks has been investigated in [9]. In [10], a multi-tenant C-RAN architecture is proposed and a

resource allocation framework is designed to perform capacity allocation between VNOs.

Nevertheless, despite these research efforts, numerous issues remain to be addressed. Most notably, RAN sharing considered in [3]–[5], [8], [9] was mainly studied under the traditional cellular architecture, which will no longer be able to support the current rapidly growing demand for broadband multimedia services. Moreover, the studies in [3]–[5], [8]–[10] have mainly focused on sharing capacity, spectrum, physical resource blocks (PRBs) and BSs whereas the computational resources in the BBU pool and the capacity of the fronthaul links have not been considered as part of the resource sharing among multiple tenants. Therefore, a comprehensive virtual resource management and network slicing framework for multi-tenant C-RANs is non-existent, this forms the motivation of our work.

In this paper, we define network slicing as a network virtualization and sharing process in which the *computational resources* of the BBU pool, the *capacity* of the fronthaul links, the *wireless radio resources* and the *physical RRHs* are shared among multiple tenants. Our objective is to design a comprehensive virtual resource management framework for small cell networks. In particular, we focus on the downlink and consider the H-CRAN architecture as the base of this study. The main contributions are summarized as follows:

1) We propose a system model of multi-tenant H-CRANs and redefine network slicing as a process of sharing computational resources of the BBU pool, the fronthaul capacity, the radio resources and the physical RRHs among the tenants.
2) We formulate the resource allocation problem as a weighted sum rate maximization problem whereby the weights corresponds to the priority of the tenants, subject to the constraints of fronthaul capacity, computational capacity of virtual machines (VMs), transmission power, user association, user admission and co-tier interference constraints.
3) We show that the resource allocation problem is a non-convex mixed-integer programming problem and propose an efficient resource allocation algorithm based on dual decomposition, whose complexity will also be analyzed.

The rest of this paper is organized as follows: Section II describes the proposed system model of the multi-tenant H-CRAN and formulates the resource allocation problem. A sub-optimal resource allocation algorithm is developed for the problem formulated in Section III. In Section IV, performance evaluation of the proposed resource allocation framework is presented and discussed. Finally, Section V concludes the paper.

II. System Model and Problem Formulation

We propose a system model for a virtualized multi-tenant H-CRAN shown in Fig. 1. In this model, we quantify the computational resources of the BBU pool as the number of VMs that can be created in the BBU pool. Each VM serves as the virtual BBU for each data flow. For simplicity, we assume that each data flow corresponds to one user. It is noteworthy that the computational resources of the BBU pool are limited. Thus, admission control of data flows becomes part of the resource sharing process. Here, we denote $\mathcal{V}$ as the set of VMs that can be supported in the BBU pool. Another assumption is made whereby all fronthaul links have equal data rate capacity. Depending on the service-level agreement (SLA) between the H-CRAN provider and the tenant, the H-CRAN may require to ensure that a minimum data rate is achieved for each user belonging to the tenant. Here, we assume that each tenant offers an equal minimum data rate for all of its associated users. The achievable data rate is also bounded by the computational capacity of the VMs. Here, we assume that the computational capacity of all the VMs is equal and is larger than the minimum data rate required by each user.

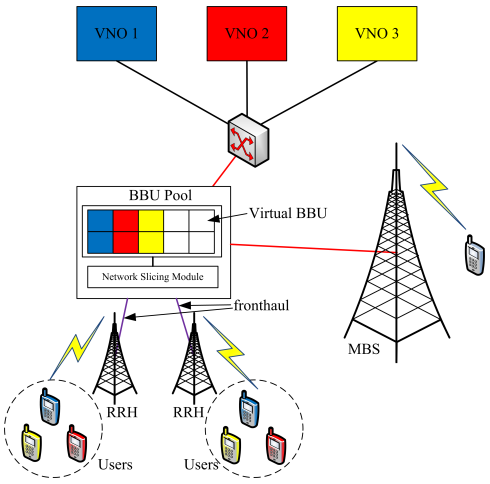

Fig. 1. Multi-tenant H-CRAN.

In our system model, the H-CRAN provider has their own users to be served, aside from providing RANaaS services to the tenants. Here, we assume that these users are served by the MBS while the users associated with the tenants are served by the small-cell RRHs. In addition, the MBS relays control signals to the tenant-associated users for resource allocation. We adopt orthogonal spectrum allocation between the macro-cell and small cells, thus alleviating cross-tier interference. Therefore, we only need to focus on network slicing in the small cell tier.

We denote $\mathcal{N}$ as the set of tenants, $\mathcal{S}$ as the set of RRHs, $\mathcal{K}$ as the set of sub-channels, $\mathcal{U}_n$ as the set of users belonging to tenant n and $\mathcal{U} = \bigcup_{n\in\mathcal{N}} \mathcal{U}_n$. We define a_u as the binary admission indicator of user u whereby $a_u = 1$ indicates that user u is admitted to the network; otherwise $a_u = 0$. Also, we define b_{su} as the binary association indicator of user u and RRH s in which $b_{su} = 1$ indicates user u associates with RRH s; otherwise $b_{su} = 0$. Further, the binary assignment indicator of sub-channel k to user u associated to RRH s is denoted by ω_{sku} whereby $\omega_{sku} = 1$ if the sub-channel is allocated to the user; otherwise $\omega_{sku} = 0$. The transmission power of RRH s

on sub-channel k for user u is denoted by p_{sku}. Besides, we define the following:

$$W(u) = w_n \qquad \forall u \in \mathcal{U}_n, \tag{1}$$

where $w_n \in (0, 1]$ is a weighting value that quantifies the priority of the users associated with tenant n.

For channel modeling, the received signal-to-interference-plus-noise ratio (SINR) of user u from RRH s on sub-channel k is given by

$$\Gamma_{sku} = \frac{g_{sku}p_{sku}}{\sum_{i\in\mathcal{S}\setminus\{s\}}\sum_{j\in\mathcal{U}\setminus\{u\}} a_j b_{ij}\omega_{ikj}g_{ikj}p_{ikj} + \sigma^2}, \tag{2}$$

where g_{sku} is the channel gain between RRH s and user u on sub-channel k and σ^2 is the additive white Gaussian noise (AWGN) power. The achievable data rate of user u associated with RRH s on sub-channel k is given by Shannon's formula as

$$R_{sku} = B\log_2(1 + \Gamma_{sku}), \tag{3}$$

where B is the bandwidth in Hz of a sub-channel. We assume that each sub-channel experiences slow and flat fading and the network is perfectly synchronized.

The network slicing problem of the virtualized multi-tenant H-CRAN can be formulated as follows:

$$\max_{\mathbf{a},\mathbf{b},\boldsymbol{\omega},\mathbf{p}} \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}} W(u)a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R_{sku}, \tag{4}$$

subject to

$$\sum_{u\in\mathcal{U}} a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R_{sku} \le R_{\text{fh}} \quad \forall s \in \mathcal{S} \tag{4a}$$

$$a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R_{sku} \le a_u b_{su} R_{\max} \quad \forall u \in \mathcal{U}, s \in \mathcal{S} \tag{4b}$$

$$a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R_{sku} \ge a_u b_{su} R_{\min,u} \quad \forall u \in \mathcal{U}, s \in \mathcal{S} \tag{4c}$$

$$\sum_{u\in\mathcal{U}} a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}p_{sku} \le P_{\max,s} \quad \forall s \in \mathcal{S} \tag{4d}$$

$$\sum_{i\in\mathcal{S}\setminus\{s\}}\sum_{u\in\mathcal{U}} a_u b_{iu}\omega_{iku}g_{iku}p_{iku} \le I_{\max} \quad \forall s \in \mathcal{S}, k \in \mathcal{K} \tag{4e}$$

$$\sum_{u\in\mathcal{U}} a_u \le |\mathcal{V}| \tag{4f}$$

$$a_u \sum_{s\in\mathcal{S}} b_{su} \le 1 \quad \forall u \in \mathcal{U} \tag{4g}$$

$$\sum_{u\in\mathcal{U}} a_u b_{su}\omega_{sku} \le 1 \quad \forall s \in \mathcal{S}, k \in \mathcal{K} \tag{4h}$$

$$p_{sku} \ge 0 \quad \forall u \in \mathcal{U}, s \in \mathcal{S}, k \in \mathcal{K} \tag{4i}$$

$$a_u, b_{su}, \omega_{sku} \in \{0, 1\} \quad \forall u \in \mathcal{U}, s \in \mathcal{S}, k \in \mathcal{K}, \tag{4j}$$

where $\mathbf{a} = \{a_1, \ldots, a_{|\mathcal{U}|}\}$, $\mathbf{b} = \{b_{11}, \ldots, b_{|\mathcal{S}||\mathcal{U}|}\}$, $\boldsymbol{\omega} = \{\omega_{111}, \ldots, \omega_{|\mathcal{S}||\mathcal{K}||\mathcal{U}|}\}$, $\mathbf{p} = \{p_{111}, \ldots, p_{|\mathcal{S}||\mathcal{K}||\mathcal{U}|}\}$, $W(u)$ is the weight function of user u associated to its tenant, R_{fh} is the fronthaul capacity, $R_{\max}$ is the computational capacity of VMs, $R_{\min,u}$ is the minimum data rate guaranteed for user u by its associated tenant, $P_{\max,s}$ is the transmission power budget of RRH s and $I_{\max}$ is the predefined tolerable interference threshold of an RRH on a sub-channel. In (4), constraint (4a) guarantees that the traffic carried by each RRH does not exceed its fronthaul capacity. Constraints (4b) and (4c) ensure that the achievable data rate of each user is bounded by the computational capacity of VMs while meeting the minimum bit rate requirement offered by the associated tenant, respectively. The transmission power of each RRH is restricted in constraint (4d). In constraint (4e), co-tier interference experienced on each sub-channel is limited by a predefined tolerable interference threshold, $I_{\max}$. The number of admitted users is ensured in constraint (4f) to be not exceeding the number of VMs available in the BBU pool, i.e., $|\mathcal{V}|$, where $|.|$ denotes the cardinality of the set. Constraint (4g) ensures that each admitted user is associated with only one RRH. Constraint (4h) enforces that no two or more users associated with the same RRH receive the same sub-channel. The transmission power of each RRH on each sub-channel is ensured to be non-negative in constraint (4i). Constraint (4j) guarantees that a_u, b_{su} and ω_{sku} take only binary values. We observe that (4) is a non-convex mixed-integer programming problem, which is generally NP-hard. An exhaustive search is not viable in this case as the computational complexity would be of exponential time. In the next section, we design an efficient resource allocation algorithm to solve (4) based on a dual decomposition method.

III. Solution Algorithm

To design an efficient resource allocation algorithm, we need to transform (4) into a tractable optimization problem. Firstly, we express the lower bound of the SINR that can be experienced by user u associated with RRH s on sub-channel k as

$$\Gamma^{\text{LB}}_{sku} = \frac{p_{sku}g_{sku}}{I_{\max} + \sigma^2}. \tag{5}$$

Subsequently, the lower bound of the achievable data rate of user u associated with RRH s on sub-channel k can be given by

$$R^{\text{LB}}_{sku} = B\log_2(1 + \Gamma^{\text{LB}}_{sku}). \tag{6}$$

Then, the lower bound of (4) can be written as

$$\max_{\mathbf{a},\mathbf{b},\boldsymbol{\omega},\mathbf{p}} \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}} W(u)a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R^{\text{LB}}_{sku}, \tag{7}$$

subject to (4d)-(4j) and

$$\sum_{u\in\mathcal{U}} a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R^{\text{LB}}_{sku} \le R_{\text{fh}} \quad \forall s \in \mathcal{S} \tag{7a}$$

$$a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R^{\text{LB}}_{sku} \le a_u b_{su} R_{\max} \quad \forall u \in \mathcal{U}, s \in \mathcal{S} \tag{7b}$$

$$a_u b_{su} \sum_{k\in\mathcal{K}} \omega_{sku}R^{\text{LB}}_{sku} \ge a_u b_{su} R_{\min,u} \quad \forall u \in \mathcal{U}, s \in \mathcal{S}. \tag{7c}$$

Although the problem in (7) can be solved directly using the dual decomposition method, it is still computationally exhaustive because many Lagrange multipliers need to be updated, thereby resulting in slow convergence. As such, we first employ a sub-optimal approach to simplify the problem by solving $\mathbf{a}$ and $\mathbf{b}$ using Algorithm 1.

Algorithm 1 User admission and association algorithm

1: Set $a_u = 0$ and $b_{su} = 0$ for all $u \in \mathcal{U}$ and $s \in \mathcal{S}$, $U_{\text{adm}} = \emptyset$.
2: Estimate the wideband SINR for all $u \in \mathcal{U}$ and $s \in \mathcal{S}$ using (8).
3: Set $b_{su} = 1$ such that the wideband SINR received by user u from RRH s is the largest among all other RRHs.
4: Estimate the weighted user throughput per sub-channel for all $u \in \mathcal{U}$ based on the received wideband SINR using (9).
5: Sort all the users in descending order of the weighted user throughput.
6: Add the first $|\mathcal{V}|$ users with the highest weighted user throughput into set U_{adm} or add all users if the total number of users is less than $|\mathcal{V}|$.
7: Set $a_u = 1$ for all $u \in \mathcal{U}_{\text{adm}}$.

In Algorithm 1, we first assume that all users are not admitted and associated with any RRH, i.e., $a_u = 0$ and $b_{su} = 0$ for all $u \in \mathcal{U}$ and $s \in \mathcal{S}$, and U_{adm} is initialized as an empty set. Subsequently, we associate each user with an RRH such that the RRH provides the largest received wideband SINR to the user. The received wideband SINR can be estimated as

$$\Gamma_{\text{wb},su} = \frac{\bar{g}_{su} P_{\text{max},s}}{\sum_{i \in \mathcal{S} \setminus \{s\}} \bar{g}_{iu} P_{\text{max},i} + \sigma^2} \quad \forall s \in \mathcal{S}, u \in \mathcal{U}, \quad (8)$$

where $\Gamma_{\text{wb},su}$ is the wideband SINR received by user u from RRH s, $\bar{g}_{su}$ is the average channel gain between RRH s and user u across the channel bandwidth. After user association, we estimate the weighted user throughput per sub-channel based on the received wideband SINR as follows:

$$R_{\text{w},u} = \sum_{s \in \mathcal{S}} W(u) b_{su} B \log_2(1 + \Gamma_{\text{wb},su}). \quad (9)$$

Then, the users are sorted in descending order of the weighted user throughput and the first $|\mathcal{V}|$ users with the highest weighted throughput values are admitted to $\mathcal{U}_{\text{adm}}$. If the total number of users is less than $|\mathcal{V}|$, all the users will be admitted to $\mathcal{U}_{\text{adm}}$.

After solving $\mathbf{a}$ and $\mathbf{b}$ with Algorithm 1, the problem in (7) can be reduced to

$$\max_{\boldsymbol{\omega}, \mathbf{p}} \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_s} W(u) \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku}, \quad (10)$$

subject to

$$\sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \le R_{\text{fh}} \quad \forall s \in \mathcal{S} \quad (10a)$$

$$\sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \le R_{\max} \quad \forall u \in \mathcal{U}_s, s \in \mathcal{S} \quad (10b)$$

$$\sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \ge R_{\min,u} \quad \forall u \in \mathcal{U}_s, s \in \mathcal{S} \quad (10c)$$

$$\sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \omega_{sku} p_{sku} \le P_{\text{max},s} \quad \forall s \in \mathcal{S} \quad (10d)$$

$$\sum_{i \in \mathcal{S} \setminus \{s\}} \sum_{u \in \mathcal{U}_i} \omega_{iku} g_{iku} p_{iku} \le I_{\max} \quad \forall s \in \mathcal{S}, k \in \mathcal{K} \quad (10e)$$

$$\sum_{u \in \mathcal{U}_s} \omega_{sku} \le 1 \quad \forall s \in \mathcal{S}, k \in \mathcal{K} \quad (10f)$$

$$p_{sku} \ge 0 \quad \forall u \in \mathcal{U}_s, s \in \mathcal{S}, k \in \mathcal{K} \quad (10g)$$

$$\omega_{sku} \in \{0, 1\} \quad \forall u \in \mathcal{U}_s, s \in \mathcal{S}, k \in \mathcal{K}, \quad (10h)$$

where $\mathcal{U}_s = \{u \in \mathcal{U} | a_u = 1, b_{su} = 1\}$. Thereafter, we assume that $\omega_{sku} = 0$ and $p_{sku} = 0$ for all $u \in \mathcal{U} \setminus \mathcal{U}_s$, $s \in \mathcal{S}$ and $k \in \mathcal{K}$, and employ the dual decomposition method to solve (10) under the assumption that Slater's condition is satisfied, i.e., the duality gap is zero. In fact, it has been proved that the duality gap for resource allocation in multi-carrier systems is nearly zero if the number of sub-channels is sufficiently large [11]. Therefore, the solution to (10) can be obtained by solving its dual problem.

We first write the Lagrangian function of (10) as

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\omega}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau}) \\
&= \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_s} W(u) \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \\
&+ \sum_{s \in \mathcal{S}} \alpha_s \left(R_{\text{fh}} - \sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \right) \\
&+ \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_s} \beta_{su} \left(R_{\max} - \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \right) \\
&+ \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_s} \phi_{su} \left(\sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} - R_{\min,u} \right) \\
&+ \sum_{s \in \mathcal{S}} \lambda_s \left(P_{\text{max},s} - \sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \omega_{sku} p_{sku} \right) \\
&+ \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \mu_{sk} \left(I_{\max} - \sum_{i \in \mathcal{S} \setminus \{s\}} \sum_{u \in \mathcal{U}_i} \omega_{iku} g_{iku} p_{iku} \right) \\
&+ \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} \tau_{sk} \left(1 - \sum_{u \in \mathcal{U}_s} \omega_{sku} \right),
\end{aligned} \quad (11)$$

where α_s, β_{su}, ϕ_{su}, λ_s, μ_{sk} and τ_{sk} are the Lagrange multipliers corresponding to constraints (10a)-(10f), respectively. Also, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_{|\mathcal{S}|}\}$, $\boldsymbol{\beta} = \{\beta_{11}, \ldots, \beta_{|\mathcal{S}||\mathcal{U}|}\}$, $\boldsymbol{\phi} = \{\phi_{11}, \ldots, \phi_{|\mathcal{S}||\mathcal{U}|}\}$, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_{|\mathcal{S}|}\}$, $\boldsymbol{\mu} = \{\mu_{11}, \ldots, \mu_{|\mathcal{S}||\mathcal{K}|}\}$ and $\boldsymbol{\tau} = \{\tau_{11}, \ldots, \tau_{|\mathcal{S}||\mathcal{K}|}\}$. It is noteworthy that (10g) and (10h) are boundary constraints which will be absorbed in the Karush-Kuhn-Tucker (KKT) conditions [12], [13] in which optimal solution is guaranteed. Therefore, the corresponding terms in (11) can be omitted. Then, the Lagrangian dual function can be expressed as

$$D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \max_{\boldsymbol{\omega}, \mathbf{p}} \mathcal{L}(\boldsymbol{\omega}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau}), \quad (12)$$

and the dual optimization problem can be formulated as

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau}} D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau}), \quad (13)$$

subject to $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau} \ge 0$.

To solve (13), we first assume that user $u \in \mathcal{U}_s$ is allocated sub-channel k, i.e., $\omega_{sku} = 1$. Taking the derivative of (11) with respect to p_{sku} yields the following KKT conditions:

$$\frac{\partial \mathcal{L}}{\partial p_{sku}} = \omega_{sku}(G_{sku} - \lambda_s) \le 0, \quad (14)$$

$$\omega_{sku} p_{sku}(G_{sku} - \lambda_s) = 0, \quad (15)$$

where

$$G_{sku} = \frac{(W(u) - \alpha_s - \beta_{su} + \phi_{su})Bg_{sku}}{(I_{\max} + \sigma^2 + p_{sku}g_{sku})\ln 2} - \sum_{i \in \mathcal{S}\setminus\{s\}} \mu_{ik} g_{sku}. \tag{16}$$

From (14)-(15), optimal power allocation can be derived as

$$p_{sku} = \left[\frac{B(W(u) - \alpha_s - \beta_{su} + \phi_{su})}{\left(\lambda_s + \sum_{i \in \mathcal{S}\setminus\{s\}} \mu_{ik} g_{sku}\right)\ln 2} - \frac{I_{\max} + \sigma^2}{g_{sku}} \right]^+ \tag{17}$$

for all $u \in \mathcal{U}_s$, $s \in \mathcal{S}$ and $k \in \mathcal{K}$, where $[x]^+ = \max(0, x)$.

Taking the derivative of (11) with respect to ω_{sku} gives the following KKT conditions:

$$\frac{\partial \mathcal{L}}{\partial \omega_{sku}} = H_{sku} - \tau_{sk} \leq 0, \tag{18}$$

$$\omega_{sku}(H_{sku} - \tau_{sk}) = 0, \tag{19}$$

where

$$H_{sku} = (W(u) - \alpha_s - \beta_{su} + \phi_{su})R^{\text{LB}}_{sku} - \left(\lambda_s + \sum_{i \in \mathcal{S}\setminus\{s\}} \mu_{ik} g_{sku}\right) p_{sku}. \tag{20}$$

From (18)-(19), it is seen that $H_{sku} - \tau_{sk} \leq 0$ if $\omega_{sku} = 0$ and $H_{sku} - \tau_{sk} = 0$ if $\omega_{sku} = 1$. In addition, one sub-channel can only be allocated to one user among those associated with the same RRH, cf. constraint (10f). Hence, given the optimal $\mathbf{p}$, optimal sub-channel allocation can be obtained as

$$\omega_{sku^*} = \begin{cases} 1 & u^* = \arg\max_{u \in \mathcal{U}_s} H_{sku} \\ 0 & \text{otherwise} \end{cases} \quad \forall s \in \mathcal{S}, k \in \mathcal{K}. \tag{21}$$

Then, upon obtaining the solution $\{\boldsymbol{\omega}, \mathbf{p}\}$, we can solve the dual problem iteratively using the sub-gradient method [14], whereby the Lagrange multipliers are updated for all $u \in \mathcal{U}_s$, $s \in \mathcal{S}$ and $k \in \mathcal{K}$, as follows:

$$\alpha_s^{(t+1)} = \left[\alpha_s^{(t)} - \delta_1 \left(R_{\text{fh}} - \sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \right) \right]^+ \tag{22a}$$

$$\beta_{su}^{(t+1)} = \left[\beta_{su}^{(t)} - \delta_2 \left(R_{\max} - \sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} \right) \right]^+ \tag{22b}$$

$$\phi_{su}^{(t+1)} = \left[\phi_{su}^{(t)} - \delta_3 \left(\sum_{k \in \mathcal{K}} \omega_{sku} R^{\text{LB}}_{sku} - R_{\min,u} \right) \right]^+ \tag{22c}$$

$$\lambda_s^{(t+1)} = \left[\lambda_s^{(t)} - \delta_4 \left(P_{\max,s} - \sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \omega_{sku} p_{sku} \right) \right]^+ \tag{22d}$$

$$\mu_{sk}^{(t+1)} = \left[\mu_{sk}^{(t)} - \delta_5 \Big(I_{\max} - \sum_{i \in \mathcal{S}\setminus\{s\}} \sum_{u \in \mathcal{U}_i} \omega_{iku} g_{iku} p_{iku} \Big) \right]^+, \tag{22e}$$

where $\alpha_s^{(t)}$, $\beta_{su}^{(t)}$, $\phi_{su}^{(t)}$, $\lambda_{su}^{(t)}$ and $\mu_{sk}^{(t)}$ are the respective α_s, β_{su}, ϕ_{su}, λ_{su} and μ_{sk} at the t-th iteration, and δ_1, δ_2, δ_3, δ_4 and δ_5 are the positive step sizes that satisfy the infinite travel conditions [14]. The process of updating the solution and the Lagrange multipliers are repeated until convergence is achieved or the predefined maximum number of iterations, $T_{\max}$ has been executed. It is noteworthy that $\boldsymbol{\tau}$ is not updated because the corresponding KKT conditions have already been fulfilled.

In the proposed algorithm, it can be observed that the solution $\{\mathbf{a}, \mathbf{b}\}$ is obtained after $|\mathcal{U}|(|\mathcal{S}| + 1)$ function evaluations. On the other hand, the sub-gradient method requires $T_{\max}$ iterations for completion and $2|\mathcal{K}|\sum_{s \in \mathcal{S}}|\mathcal{U}_s|$ function evaluations for each update of the Lagrange multipliers. Thus, the computational complexity of the proposed algorithm is of $\mathcal{O}\left(|\mathcal{U}|(|\mathcal{S}| + 1) + 2T_{\max}|\mathcal{K}|\sum_{s \in \mathcal{S}}|\mathcal{U}_s|\right)$.

IV. Results and Discussion

In this section, we present some numerical results of the proposed network slicing framework. Here, we consider a two-tenant H-CRAN, which consists of a macrocell, with a radius of 500 m overlaid by ten pico-RRHs, which are equidistant between each other within the cell and are located at a distance of 250 m from the MBS. The number of sub-channels is set to 100 with each having a bandwidth of 180 kHz [15]. We investigate the throughput performance of the network by varying the number of users associated with the first tenant, denoted as VNO 1, while fixing the number of users associated to the second tenant, denoted as VNO 2, as 50. Other network parameters are set as follows: $R_{\text{fh}} = 1$ Mb/s, $R_{\max} = 200$ kb/s, $R_{\min,u} = 128$ kb/s for all $u \in \mathcal{U}$, $P_{\max,s} = 30$ dBm for $s \in \mathcal{S}$ and $|\mathcal{V}| = 80$. For channel modeling, we consider independently and identically distributed (i.i.d.) Rayleigh fading with zero mean and unit variance. We also consider log-normal shadowing which is also i.i.d. with zero mean and a standard deviation of 10 dB, and the path loss model: $140.7 + 36.7 \log d$ where d is the distance between the RRH and the user in km [16]. The noise power spectral density and noise figure are respectively set to -174 dBm/Hz and 9 dB [16]. The users are uniformly distributed within the network. For the proposed scheme, δ_1, δ_2, δ_3, δ_4 and δ_5 are set following the non-summable diminishing rule [14] and $T_{\max} = 100$. All results are averaged over 100 simulation runs. Hereafter, we first investigate and compare the network slicing performance of the proposed scheme with a baseline resource allocation scheme, which randomly admits users and treats all tenants with equal priority. The latter scheme is similar to those used for single-operator cellular networks. Then, the effect of $I_{\max}$ to the network performance of the proposed scheme is investigated.

Fig. 2 illustrates the throughput performance of a two-tenant H-CRAN with different priority weighting values when $I_{\max} = -100$ dBm. We notice that the throughput achieved by VNO 1 gradually increases with a larger number of its associated users while the throughput achieved by VNO 2 gradually reduces. This is due to the fact that the number of users associated with VNO 1 increases, which eventually exceeds that associated with VNO 2. Therefore, more users

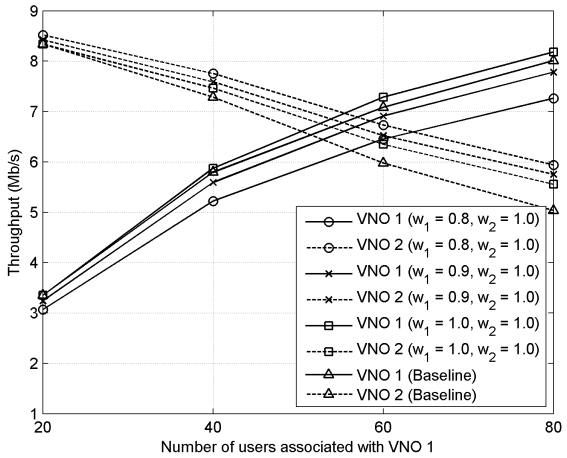

Fig. 2. Throughput performance of a two-tenant H-CRAN with dfferent priority weights.

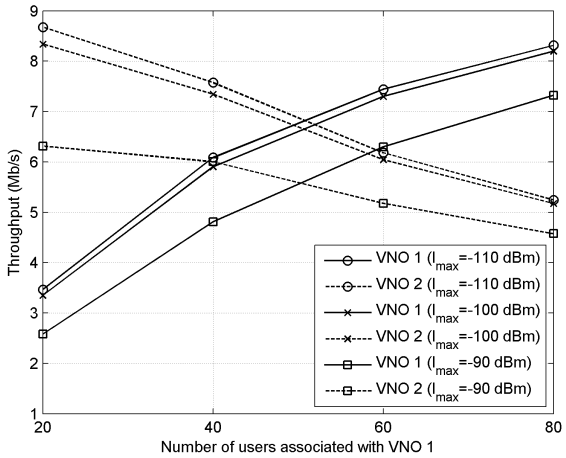

Fig. 3. Throughput performance of a two-tenant H-CRAN with different values of $I_{\max}$.

associated with VNO 1 are admitted to the network. More importantly, the throughput performance of VNO 1 becomes lower with a lower w_1 while the throughput performance of VNO 2 becomes higher[1]. This is because a low w_1 gives a lower probability of admitting the users associated with VNO 1, as evident in (9), and reduces the achievable throughput of VNO 1 when maximizing (10) while increasing the achievable throughput of VNO 2. Thus, the priority weight is an important parameter that defines the "size" of the network slice provided to each tenant, i.e., the RRHs and the amounts of computational resources of the BBU pool, fronthaul capacity and radio resources which are allocated to each tenant. On the other hand, it can be observed that the proposed scheme with $w_1 = 1.0$ and $w_2 = 1.0$ outperforms the baseline scheme, even though both schemes equally prioritize all the tenants. This thanks to Algorithm 1 in the proposed scheme, which greedily maximizes the throughput by admitting the user with the highest achievable weighted throughput. Additionally, the baseline scheme cannot differentiate the priority of the tenants, thus it will result in the same performance as in Fig. 2 regardless of the priority of the tenants.

Next, the effects of the interference constraint in (4e) to the multi-tenant H-CRAN is investigated. Fig. 3 shows the throughput performance of the two-tenant H-CRAN with different values of $I_{\max}$, and $w_1 = 1$ and $w_2 = 1$. It can be observed that the throughput performance of the VNOs improves with a lower value of $I_{\max}$ as shown in Fig. 3 where the throughput gain of the network with $I_{\max} = -100$ dBm over that with $I_{\max} = -90$ dBm is significant. This is because the low value of $I_{\max}$ allows for increasing the transmission power of the RRHs on each subchannel (cf. (17)), hence the higher throughput gain. However, the throughput gain gradually reduces with an even lower $I_{\max}$, as observed in Fig. 3 where the throughput gain of the network with $I_{\max} = -110$ dBm over that with $I_{\max} = -100$ dBm is small. This is because the allowable interference level is high enough to offset the throughput gain. Hence, the value of $I_{\max}$ needs to be properly tuned such that the network slice provided to each tenant can achieve optimal network performance.

[1]Refer to (1) for the definition of the priority weighting value of a tenant.

V. Conclusion and Future Work

In this paper, we have proposed a new network slicing framework for multi-tenant H-CRANs. In particular, we defined that network slicing as a process of sharing computational resources in the BBU pools, fronthaul capacity, physical RRHs and radio resources. We have formulated the framework as a weighted throughput maximization problem and solved it jointly using a greedy sub-optimal approach and a dual decomposition method. Numerical results show that the user priority weights associated with a particular tenant is a critical parameter that scales the network slice provided to the tenant. Also, the allowable interference level is another important parameter which optimizes the throughput performance of the multi-tenant H-CRANs. For future work, we will investigate the effect of the varying minimum user bit rate and fronthaul capacity to the network throughput.

Appendix A
List of Notation

A list of notation used in this paper is given in Table I.

Acknowledgment

This work is supported in part by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme MMUE/140082.

References

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, Mar. 2015.

TABLE I
LIST OF NOTATION.

Notation	Description
$\mathcal{V}$	Set of VMs
$\mathcal{N}$	Set of tenants
$\mathcal{S}$	Set of RRHs
$\mathcal{K}$	Set of sub-channels
$\mathcal{U}_n$	Set of users associated with tenant n
a_u	Binary admission indicator of user u
b_{su}	Binary association indicator of user u and RRH s
ω_{sku}	Binary assignment indicator of sub-channel k to user u associated with RRH s
p_{sku}	Transmission power of RRH s on sub-channel k for user u
w_n	Weighting value that quantifies the priority of the users associated with tenant n
Γ_{sku}	SINR received by user u from RRH s on sub-channel k
g_{sku}	Channel gain between RRH s and user u on sub-channel k
σ^2	AWGN power
B	Sub-channel bandwidth
R_{fh}	Fronthaul capacity
R_{max}	Computational capacity of VMs
$R_{\text{min},s}$	Minimum data rate guaranteed for user u
$P_{\text{max},s}$	Transmission power budget of RRH s
I_{max}	Predefined tolerable interference threshold of an RRH on a sub-channel
Γ^{LB}_{sku}	Lower bound of the SINR experienced by user u associated with RRH s on sub-channel k
R^{LB}_{sku}	Lower bound of the achievable data rate of user u associated with RRH s on sub-channel k
Γ_{su}	Wideband SINR received by user u from RRH s
$\bar{g}_{su}$	Average channel gain between RRH s and user u
$R_{\text{w},u}$	Weighted user throughput per sub-channel based on the received wideband SINR
$\mathcal{U}_{\text{adm}}$	Set of admitted users
$\mathcal{U}_s$	Set of users whereby $a_u = 1$ and $b_{su} = 1$
$\alpha_s, \beta_{su}, \phi_{su}, \lambda_s, \mu_{sk}, \tau_{sk}$	Lagrange multipliers corresponding to constraints (10a)-(10f)
$\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$	Positive step sizes that satisfy the infinite travel conditions for updates of α_s, β_{su}, ϕ_{su}, λ_s and μ_{sk}
t	Iteration index
T_{max}	Maximum number of iterations
d	Distance between the RRH and the user in km

[2] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiency," *IEEE Wireless Commun. Mag.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[3] F. Fu and U. C. Kozat, "Wireless network virtualization as a sequential auction game," in *Proc. INFOCOM*, San Diego, California, Mar. 2010, pp. 1–9.

[4] R. Kokku, R. Mahindra, H. Zhang, and R. S, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[5] J. S. Panchal, R. D. Yates, and M. M. Buddhikot, "Mobile network resource sharing options: performance comparisons," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4470–4482, Sep. 2013.

[6] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2015.

[7] C. Liang and F. R. Yu, "Wireless virtualization for next generation mobile cellular networks," *IEEE Wireless Commun. Mag.*, vol. 22, no. 1, pp. 61–69, Feb. 2015.

[8] P. C. Farces, X. C. Perez, K. Samdanis, and A. Branchs, "RMSC: a Cell slicing controller for virtualized multi-tenant mobile networks," in *Proc. IEEE 81st VTC Spring*, Glasgow, United Kingdom, May 2015, pp. 1–6.

[9] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN virtualization in multi-tenant LTE-A heterogeneous networks," *IEEE Trans. Veh. Technol.*, pp. 1–14, Sep. 2015, in press.

[10] S. Khatibi and L. M. Carreia, "A model for virtual radio resource management in virtual RANs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 68, pp. 1–12, Mar. 2015.

[11] W. Yu and R. Lui, "Dual Methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[12] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[13] K. Kim, Y. Han, and S.-L. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems," *IEEE Commun. Lett.*, vol. 9, no. 6, pp. 526–528, Jun. 2005.

[14] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Notes for EE364b, Stanford University, 2006-07.

[15] "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," TS 36.211, 3rd Generation Partnership Project (3GPP), Sophia-Antipolis, France, Sep. 2015. [Online]. Available: http://www.3gpp.org/DynaReport/36.211.htm

[16] "3rd generation partnership project; technical specification group tadio access network; small cell enhancements for e-utra and e-utran - physical layer aspects (release 12)," TR 36.872, 3rd Generation Partnership Project (3GPP), Sophia-Antipolis, France, Sep. 2013. [Online]. Available: http://www.3gpp.org/dynareport/36872.htm